THE UNIVERSITY OF NEW MEXICO

ECE 538
Advanced Computer Architecture

Instructor: Lei Yang

Department of Electrical and Computer Engineering

November 29, 2021

WHAT IS COMPUTER ARCHITECTURE?

- Techniques for boosting performance:
  - Instruction-Level parallelism (ILP)
  - Data-level parallelism (DLP)
  - Thread-level parallelism (TLP)

2021-11-24: End Class 27 here

- ***

Custom and Emerging Computing Trends

The Future of Computing

Heterogeneous Architecture – CPUs, GPUs, FPGAs, ASICs, ...

OVERVIEW

- Specific architectures:
  - Examples: ASIP, FPGA, and ASIC
  - Comparison of general-purpose processors: DSP, GPU
- Heterogeneous multicore platforms
  - Introduction to domain specific computing

General Purpose Processor (GPP)

- Hardware features:
  - The program to be run and necessary data could be made available at main memory
  - General data-path: consists of a general ALU and usually a large register file
  - Multiple levels of cache for reducing memory latency
- Maximum flexibility
  - Programmable: support several high-level languages
  - Can be used for any general applications
- Other key features…
  - Short time-to-market
  - Low non-recurring engineering (NRE) cost
  - High power consumption

Application Specific Instruction Set Processor (ASIP)

- A microprocessor tailed to benefit a specific application or a domain of applications (operations)

  -- Signal processing, image processing, video processing, digital communication, etc.

- Instruction set of ASIP is customized for the type of computation involved in specific application
- Requirement on flexibility should be sufficient instead of unlimited like GPP
- Highest performance
- Lower power consumption
- Smaller silicon, design cost

Digital Signal Processor (DSP)

- A Digital Signal Processor (DSP) is an ASIP designed for repetitive multiply accumulate operation and bit-reversal addressing

- Real-time processing of computation-intensive tasks while minimizing:
  - Cost (than GPP)
  - Power consumption
  - Specialized datapath
  - Specialized instruction set for DSP

- Repetitive numeric calculations
- High memory bandwidth
- Support streaming data
- Less development time

Graphic Processing Unit (GPU)

- GPU is a processor – Specialized for graphics
- Modern GPUs are very efficient at manipulating computer graphics and image processing, and their highly parallel structure makes them more efficient than general-purpose CPUs for algorithms where the processing of large blocks of data is done in parallel
- Heterogeneous execution model (CPU is the host, GPU is the device)

- The motivation of GPGPU computing:

-- Parallel computing by massively data-parallel stream processing
-- Use thousands of threads; Implementation in hundreds of cores
-- Cost effective: cheap
-- Using already available hardware

- GPUs involve high latency, high throughput processing
- CPU performs low latency low throughput computation

Field Programmable Gate Array (FPGA)

- Core architecture of FPGA consists of three main components:
  1. An array of configurable/ programmable logic blocks:
     -- combinational units
  2. A sea of programmable interconnects
  3. Memories and specialised I/O blocks

- Other components in modern FPGAs:
  - ✓ DSP blocks
  - ✓ Hardwired IP cores
  - ✓ Soft cores
  - ✓ …

FPGA vs. DSP

1. Both FPGA and DSP offer flexibility for reuse and programmability

2. DSP has an advantage in time-in-market

3. FPGA supports parallel design and offers the greater performance

4. FPGA is more expensive and power-consuming

Heterogenous arch. With DSP and FPGA components are an emerging trend

Application Specific Integrated Circuits (ASIC)

- Provide customized tailed HW solutions for specific apps. or problems

- Not configurable
- Cannot be upgraded, updates require redesign
- Involve high Non-recurring engineering (NRE) cost of several millions for development and testing (need to have large volume market to reduce cost)
- Longer design/development time (time-to-market)
- Maximum performance per watt, per unit area
- Possible design optimizations
- Semi-custom design: **gate array** implementation, **standard array** implementation
- Full-custom design: cell design, cell library development and use, simulation and testing for design verification

GATE ARRAYS AND STANDARD CELLS

FPGA vs. ASIC

- Differences
- FPGA:
  - ✓ Reconfigurable, shorter development time
  - ❖ more power and energy, expensive for large volume market
- ASIC:
  - ✓ higher clock Freq, higher speed, smaller
  - ❖ Development tools are expensive, not flexible

- Integration benefits:
  - ✓ both have highly parallel tasks to provide greater performance,

CPU vs. GPU vs. FPGA vs. ASIC

CPU vs. GPU vs. FPGA vs. ASIC

	CPU	GPU	FPGA	ASIC
Compute Adaptability (to a variety of situations)	High	Medium	Low	None
Compute power	Medium	High	High	Medium
Latency	Medium	High	Low	Ultra low
Throughput	Low	High	High	High
Parallelism	Low	High	High	High
Power efficiency	Medium	Low	Medium	High

COMPARISON

ASIC	Application Specific Integrated Circuit
CGRA	Coarse-Grained Reconfigurable Architecture
FPGA	Field-Programmable Gate Array
ASIP	Application-Specific Instruction Set Processor
GPP	General Purpose Processor

THE UNIVERSITY OF NEW MEXICO

ECE 538
Advanced Computer Architecture

Instructor: Lei Yang

Department of Electrical and Computer Engineering

November 29, 2021

2021-11-24: End Class 27 here

- CPU
- GPU
- FPGA
- ASIC

ASIC	Application Specific Integrated Circuit
CGRA	Coarse-Grained Reconfigurable Architecture
FPGA	Field-Programmable Gate Array
ASIP	Application-Specific Instruction Set Processor
GPP	General Purpose Processor

The Future of Computing

Heterogeneous Architecture – CPUs, GPUs, FPGAs, ASICs, ...

HETEROGENEOUS MULTICORE PLATFORMS

- ARM big. LITTLE

  -- Selects processor according to job to minimise power consumption

  -- Cores need not be always on

Two different types of cores:

- big: ARM Cortex-A15: consists of out-of-order multi-issue pipeline high-performance CPUs

  -- Does the computation-intensive tasks

- LITTLE: ARM Cortex-A7: consists of in-order, 8-stage pipeline CPUs

  -- Perform tasks which demand less computation

HETEROGENEOUS MULTICORE PLATFORMS

IBM Cell Broadband Engine (BE) architecture

- Two types of cores: customised for gaming/graphics rendering.
- Sony-Toshiba-IBM (STI) partnership built CELL processor was used in Sony's PlayStation 3 (2000-2005).

PPE: assign tasks to other cores; control communication between cores

SPE: perform vector operations; provide high floating-point throughput

General Heterogeneous Computing Systems

- Heterogeneous systems can combine different processing technologies: FPGA, ASIC, GPP, GPU, DSP or microcontroller units to achieve desired performance with less power consumption.

.....

DOMAIN SPECIFIC COMPUTING (DSC)

- A domain specific computing system is a heterogeneous multicore system

- It is customised specifically according to the computations (the kind of processing) involved in a given application domain for high-performance and power efficient realisation of the computations for the particular application domain and support the changes in the application domain.

- A DSC system may contain some generic programmable core, customised cores, hardware accelerators and programmable interconnects.

Domain Specific Computing (DSC)

- Motivation:
  - Each user has a high computing demand in one or a few selected application domains
  - Customizable computing systems for particular app. domain for perf/Power efficiency
  - Make best use of techniques (ASIC, FPGA…) for power/perf efficiency with lower cost

- E.g., **healthcare application**, **graphics** for game developers, **circuit simulation** for integrated circuit design houses, **financial analytics** for investment banks, **transforms and filters** for signal/image and video processing, **matrix operations** for weather forecasting, scientific computing

- General computing needs like email, word processing, web browsing can be easily achieved by any existing computing technology

Domain Specific Computing (DSC)

- Example 1

- Low-latency, high-bandwidth interconnects for data sharing between cores, co-processors, cache banks and memory banks to meet communication requirements of a particular application (or even different phases of the same application).

Domain Specific Computing (DSC)

Example

- Low-latency, high-bandwidth interconnects for data sharing between cores, co-processors, cache banks and memory banks to meet communication requirements of a particular application (or even different phases of the same application).
- CHP is characterised by three types of components (fixed cores, customisable cores, and programmable fabrics); each has different levels of customisation, re-configurability and parallelism.

Domain Specific Computing (DSC)

- Customizable heterogeneous platform of a DSC system
  - Fixed cores: differ widely in terms of energy efficiency, performance and area

    -- They have very limited (or no) re-configurability

    -- Allow voltage/frequency scaling to adapt power/performance
  - Customizable cores: architecture customised (tuned) according to application

    -- Offer a set of tunable options (e.g., register file sizes, cache sizes, bit-width of datapath operating frequency and supply voltages)
  - Programmable fabrics: provide maximum flexibility

    -- Implement custom instructions by specialised co-processor (FFT, DCT, filters, etc.) to accelerate performance

    -- Architecture customised in terms of # computing units, types of computing units, # pipeline stages, etc. to implement complex operations

Processor Technology Summary

- Programmable Processors
  - General purpose processor
    - -- Single-core and multicore processors
  - Specialised programmable processor
    - -- Application Specific Instruction Set Processors (ASIP), GPU, DSP
- Dedicated Processors
  - Reconfigurable processors (e.g., FPGA)
  - Application Specific Integrated Circuits (ASIC)
- Heterogeneous System On Chip (SoC)
  - ARM big. LITTLE
  - IBM Cell Broadband Engine (BE) architecture

Choice of Implementation Platforms

- System cost; Speed; Power consumption; Design reuse; Time-to-Market

THE FUTURE OF COMPUTING AND ARCHITECTURES

(Click Here)

WHAT DID WE LEARN SO FAR

- Applications of specific architectures:
  - ✓ Examples: ASIP, FPGA, and ASIC
  - ✓ Comparison of general purpose processors: DSP, GPU, FPGA and ASIC
- Heterogeneous multicore platforms
  - ✓ Domain Specific Computing Systems

2021-11-29: End Class 28 here